Bayes Linear Analysis for Complex Physical Systems Modeled by Computer Simulators

Michael Goldstein [⋆]

Department of Mathematical Sciences,
Durham University
Durham, United Kingdom

Abstract. Most large and complex physical systems are studied by mathematical models, implemented as high dimensional computer simulators. While all such cases differ in physical description, each analysis of a physical system based on a computer simulator involves the same underlying sources of uncertainty. These sources are defined and described below. In addition, there is a growing field of study which aims to quantify and synthesize all of the uncertainties involved in relating models to physical systems, within the framework of Bayesian statistics, and to use the resultant uncertainty specification to address problems of forecasting and decision making based on the application of these methods. We present an overview of the current status and future challenges in this emerging methodology, illustrating with examples drawn from current areas of application including: asset management for oil reservoirs, galaxy modeling, and rapid climate change.

Keywords: Bayes linear analysis, computer experiment, emulation, history matching, forecasting, reification

1 Uncertainty in complex systems represented by computer simulators

Most large and complex physical systems are studied by mathematical models, implemented as high dimensional computer simulators (like climate models). To use complex simulators to make statements about physical systems (like climate), we need to quantify the uncertainty involved in moving from the model to the system. The issues that we must address are methodological (how can we estimate what climate is likely to be?), computational (how can we ensure that our methods are tractable?) and foundational (why should our methods work and what do our answers mean?).

[⋆] Thanks to Engineering and Physical Sciences Research Council for funding through the Managing Uncertainty in Complex models project, and Natural Environment Research Council for funding through the RAPID programme. Thanks to Jonathan Cumming, Ian Vernon, Danny Williamson for the collaborations described in this paper.

Applications range across all areas of science and technology. In this article, I will offer illustrations based on the following three applications, chosen as I have some personal experience with each, and they illustrate the wide range of areas of application for the methodology that we shall describe.

Oil reservoirs An oil reservoir simulator takes as inputs a physical description of the properties of a reservoir (permeabilities, porosities, faults, etc) and produces, as output, various well characteristics(pressure profiles, oil and gas production rates, etc.). The simulator is used to help manage assets associated with the reservoir. The aim is commercial: to develop efficient production schedules, determine whether and where to sink new wells, and so forth.

Galaxy formation The study of the development of the Universe is supported by using Galaxy formation simulators. These simulators take as input various parameters controlling physical processes which are thought to control the formation of galaxies and the simulators perform simulations of the development of the universe from the point of origin to the present producing as output various large scale quantities which can be compared to current cosmological measurements. The aim is scientific - to gain information about the physical processes underlying the Universe.

Climate change Large scale climate simulators are constructed to assess likely effects of human intervention upon future climate behaviour. Aims are both scientific - much is unknown about the large scale interactions which determine climate - and also very practical, as such simulators provide evidence for the importance of changing human behaviour before possibly irreversible changes are set into motion.

While each such model differs in all details of the scientific basis and mathematical implementation, there are various sources of uncertainty which are common across all such applications.

(i) parametric uncertainty (each model requires a, typically high dimensional, parametric specification, whose value is not known),

(ii) condition uncertainty (uncertainty as to boundary conditions, initial conditions, and forcing functions),

(iii) functional uncertainty (model evaluations often take a long time, so the function is unknown for almost all choices of inputs),

(iv) stochastic uncertainty (either the model is stochastic, giving different outcomes each time it is evaluated under the same choice of input parameters, introducing uncertainly directly, or aspects of the modelling which should involve such stochastic uncertainty have been reduced to a deterministic form, introducing uncertainty indirectly),

(v) solution uncertainty (the system equations can only be solved to some necessary level of approximation),

(vi) structural uncertainty (even taking into account all of the above sources of uncertainty, the model only approximates the physical system and this discrepancy introduces further uncertainty about system behaviour),

(vii) measurement uncertainty (as the model is calibrated against system data all of which is measured with error),

(viii) multi-model uncertainty (usually we have not one but many models related to the physical system),

(ix) decision uncertainty (to use the model to influence real world outcomes, we need to relate things in the world that we can influence to inputs to the simulator and through outputs to actual impacts. These links are uncertain).

Different physical models vary in many aspects, but the formal structures for analyzing all of these components of the uncertainty for the physical system, as derived from the study of computer simulators for the system, are very similar, which is why there is a common underlying methodology for such problems. In this article, we give an informal introduction to some important features of this methodology. First, we introduce some general features of the uncertainty structure, describe the Bayesian approach for addressing these uncertainties, and explain why we prefer, in certain cases, a Bayes linear approach to the uncertainty analysis. Then, we outline the rationale for history matching as a way to constrain the input space, and describe a simple forecasting methodology for future system outcomes. We illustrate the development with a brief description of examples arising from each of the simulation problems described above. Finally, we consider the reason why we should view a computer simulator as being informative for a physical system.

2 General uncertainty structure

Each simulator for a physical system can be conceived as a function $F(x)$, where x is an input vector, representing unknown properties of the physical system, and $F(x)$ is the corresponding output vector representing aspects of system behaviour, y.

Interest in the analysis concerns general qualitative insights as to the behaviour of the system plus some of the following: (i) the "appropriate" (in some sense) choice, x^*, for the system properties x; (ii) the use that we can make of historical observations z, observed with error on a subset y_h of y, both to test and to constrain the model; (iii) how informative $F(x^*)$ is for actual system behaviour, y, particularly for forecasting future system outcomes, y_p;(iv) the optimal assignment of any decision inputs, d, in the model.

For example, in a climate analysis, y_h might correspond to historical climate outcomes over space and time, y to past, current and future climate, and the "decisions" might correspond to different policy relevant choices such as carbon emission scenarios.

How can we solve such problems? If observations, z, are made without error and the model is a perfect reproduction of the system, then, in principle, we can write $z = F_h(x^*)$, invert f_h to find x^* and learn about all future components of $y = F(x^*)$. If x contains some control parameters, then these are set to optimize properties of future outcomes contained in y.

However, in practice, the inversion of slow, high dimensional and complex functions is a very hard problem. Further, the observations z are typically made with error, and the model always differs from the physical system, so we must separate the uncertainty representation into two relations, one expressing the data uncertainty and one expressing the structural uncertainty: for example, the simplest such representation is of the form

$$z = y_h \oplus e, \; y = F(x^*) \oplus \epsilon \tag{1}$$

where e, ϵ have some appropriate probabilistic specification, possibly involving parameters which require estimation, and the notation $U \oplus V$ denotes the addition of $U + V$, in the case where U is probabilistically independent of V, for a full probabilistic specification or U, V are uncorrelated, if we only make a second order specification of means, variances and covariances. We therefore need to make a statistical inversion of the data through the function and make statistical predictions as to future system behaviour. This is a much harder problem than the deterministic inversion, and we still haven't accounted for condition uncertainty, multi-model uncertainty, and so forth.

In practice it is extremely rare to find a serious quantification of the total uncertainty about a complex system arising from the all of the uncertainties in the model analysis that we have identified. Therefore, for almost all applications, no-one really knows the reliability of the model based analysis, so that there can be no sound basis for identifying appropriate real world decisions based on such analyses. The space between models and reality arises partly because modellers and scientists don't think about total uncertainty in a sufficiently systematic way and nor do most statisticians. Policy makers don't know how to frame the right questions for the modellers to identify the gap between their analyses and the likely outcomes in the real world and there are few funding mechanisms to address such issues. And, of course , such a full uncertainty analysis is difficult and time consuming.

3 Bayesian uncertainty analysis for complex models

In the subjectivist Bayesian view, the meaning of any probability statement is the uncertainty judgement of a specified individual, expressed on the scale of probability (by consideration of some operational elicitation scheme, for example by consideration of betting preferences); for a somewhat subjective introduction to the subjectivist position, see[6]. This interpretation has an agreed testable meaning, sufficiently precise to act as the basis of a discussion about the meaning of the analysis. In this interpretation, any probability statement is the judgement of a named individual, so we should speak not of the probability of rapid climate change, but instead of Anne's probability or Bob's probability of rapid climate change and so forth.

There is an important practical issue of perception, as most people expect something more authoritative and objective than a probability which is one person's judgement. However, the disappointing truth is that, in almost all cases, stated probabilities emerging from a complex analysis are not even the judgements of any individual. Nor do they have any other clear and well defined meaning.

So, it is not unreasonable that the objective of our analysis should be probabilities which are asserted by at least one person (more would be good!). The Bayesian formalism provides a way, at least in principle, to realize this aim. In the simplest form, the Bayesian approach requires the specification of the following ingredients:

– a prior probability distribution for best inputs x^*
– a probabilistic uncertainty description for the computer function F
– a probabilistic discrepancy measure relating $F(x^*)$ to the system y
– a likelihood function relating historical data z to y

This full probabilistic description provides a formal framework to synthesis expert elicitation, historical data and a careful choice of simulator runs. We may then use our collection of computer evaluations and historical observations to analyze the physical process in order to determine appropriate values for simulator inputs (calibration; history matching), to assess the future behaviour of the system (forecasting), and to optimize the performance of the system.

There is much current interest in this problem. Good starting points for entering the Bayesian literature in this area are [10], [12]. A great general resource, offering references, papers, discussion and a methodological toolkit, is the Managing Uncertainty in Complex Models (MUCM) web-site, `http://www.mucm.ac.uk/`. (MUCM is a consortium between the Universities of Aston, Durham, LSE, Sheffield, Southampton, developing general methodology for this general area with Basic Technology funding.)

This approach is very successful for problems of intermediate size and complexity. For very large scale problems, however, such a full Bayes analysis is very difficult because (i) it is hard to give a meaningful full prior probability specification over high dimensional spaces; (ii) the computations for learning from data (observations and computer runs), particularly for identifying informative ensembles of choices of parameter values at which to evaluate the simulator, may be technically difficult; (iii) the likelihood surface is extremely complicated, and any full Bayes calculation may therefore be extremely non-robust.

4 Bayes linear approach

The idea of the Bayesian approach, namely capturing our expert prior judgements in stochastic form and modifying them by appropriate rules given observations, is conceptually appropriate (and there is no obvious alternative). Bayes linear analysis is a practical alternative to the fully specified Bayesian approach, being based on a prior specification only of the means, variances and covariances of all quantities of interest, where we make expectation, rather than probability, the primitive for the theory, following de Finetti, [5]. For a full account of the Bayes linear approach, see [9].de Finetti chooses expectation over probability as, if expectation is primitive, then we can choose to make as many or as few expectation statements as we choose (including our choice of probabilities, which are simply expectations for the corresponding indicator functions), whereas, if probability is primitive, then we must make all of the probability statements before we can make any of the expectation statements. When there are many quantities that we must specify uncertainty judgements for, it is very helpful to have the option of restricting our attention to that sub-collection of specifications which we are most interested in analyzing carefully.

Corresponding to Bayes theorem, which is the basic updating tool for full Bayes analysis, is the operation of belief adjustment. $\mathsf{E}_z[y], \mathsf{Var}_z[y]$ are the expectation and variance for the vector y adjusted by the vector z, evaluated as

$$\begin{aligned}\mathsf{E}_z[y] &= \mathrm{E}(y) + \mathrm{Cov}(y,z)\mathrm{Var}(z)^{-1}(z - \mathrm{E}(z)),\\ \mathsf{Var}_z[y] &= \mathrm{Var}(y) - \mathrm{Cov}(y,z)\mathrm{Var}(z)^{-1}\mathrm{Cov}(z,y).\end{aligned}$$

If $\mathrm{Var}(z)$ is not invertible, then we use an appropriate generalized inverse.

Bayes linear adjustment may be viewed as an approximation to a full Bayes analysis or the appropriate analysis given a partial specification based on expectation as primitive. The foundation for the approach is an explicit treatment of temporal uncertainty, and the underpinning mathematical structure is the inner product space, as opposed to the probability space, which is simply a special case. The adjusted expectation of y given z is the linear combination of the elements of z, plus the unit constant, which minimizes the expected squared distance to y. Observe that de Finetti's primitive definition for conditional expectation (see [5]) corresponds to this definition in the special case in which the vector $z = (z_1, \ldots, z_r)$ represents the elements of a partition (so that one and only one of the elements of z will equal 1, and all other elements will equal 0). In this special case, adjusted expectation is equivalent to conditional expectation, so that the definition of conditioning may be viewed as a special case of that for belief adjustment, in which the vector z is restricted to a partition vector. There are other special cases in which adjusted expectation and conditional expectation coincide, the most important being that of the multivariate Gaussian distribution.

Full Bayes analysis can be more informative than the Bayes linear counterpart, if done extremely carefully, both in terms of the prior specification and the analysis. Bayes linear analysis is partial but easier, faster, and often more robust particularly for history matching and forecasting. The examples discussed below were all carried out within the Bayes linear approach. However, the ideas and approaches are complementary and there are natural full Bayes counterparts for each of the analyses that we describe.

5 Function emulation

Uncertainty analysis, for high dimensional problems, is even more challenging if the function $F(x)$ is expensive, in time and computational resources, to evaluate for any choice of x. For example, large climate models can take many weeks to evaluate on extremely powerful computers.

In such cases, $F(x)$ must be treated as uncertain for all input choices except the small subset for which an actual evaluation has been made. Therefore, we must construct a description of the uncertainty about the value of $F(x)$ for each possible choice of x. Such a representation is often termed an emulator of the function - the emulator both suggests an approximation to the function and also contains an assessment of the likely magnitude of the error of the approximation. We use the emulator either to provide a full joint probabilistic description of all of the function values (full Bayes) or to assess expectations variances and covariances for pairs of function values (Bayes linear).

There are many ways to construct emulators for computer models. A good introduction to this area is [11]. A common choice of form for the emulator is as follows. We express the emulator for component F_i of F as

$$f_i(x) = \sum_j \beta_{ij} g_{ij}(x) \oplus u_i(x)$$

where $B = \{\beta_{ij}\}$ are unknown scalars, g_{ij} are known deterministic functions of x,$u_i(x)$ is a weakly second order stationary stochastic process. There are many choices of correlation function for this process; a common choice is

$$\text{Corr}(u_i(x), u_i(x')) = \exp\left(-\left(\frac{\|x - x'\|}{\theta_i}\right)^2\right)$$

In this representation, $Bg(x)$ expresses global variation, i.e. aspects of the overall behaviour of the function that we can discover from a design which is well dispersed in parameter space, while $u(x)$ expresses local variation, i.e. those aspects of the behaviour of the function which can only be assessed by making function evaluations in the neighbourhood of x.

We fit the emulators, given a collection model evaluations, using our favourite statistical tools, such as generalized least squares, maximum likelihood, full Bayes or Bayes linear, aided wherever possible by detailed expert judgement. We need careful (multi-output) experimental design to choose informative model evaluations, and detailed diagnostics to check emulator validity.

If the simulator is really slow to evaluate, then a practical way to develop the emulator is to model jointly the simulator with a fast approximate version, F'. So, for example, based on many fast simulator evaluations, we build emulator

$$f'_i(x) = \sum_j \beta'_{ij} g_{ij}(x) \oplus u'_i(x)$$

We use this form as the prior specification for the emulator $f_i(x)$. Then a relatively small number of evaluations of $F_i(x)$, combined with relations such as

$$\beta_{ij} = \alpha_i \beta'_{ij} + \gamma_{ij}$$

enables us to adjust the prior emulator to an appropriate posterior emulator for $F_i(x)$. This approach exploits the heuristic that we need many more function evaluations to identify the qualitative form of the model (i.e. choose appropriate forms $g_{ij}(x)$, etc) than to assess the quantitative form of all of the terms in the model - particularly if we fit meaningful regression components to account for a large component of global variation.

6 History matching

Model calibration aims to identify "true" input parameters x^*. However full Bayes calibration analysis may be technically difficult and non-robust. Further, we may not believe in a unique true input value for the model and, indeed, we may be unsure whether there are any good choices of input parameters (due to model deficiencies).

A conceptually simple alternative, or precursor, to calibration is "history matching", i.e. finding the collection of all input choices x for which we judge the match of the model to the data to be acceptable, using some 'implausibility measure' $I(x)$ based on a natural probabilistic metric, accounting for emulator uncertainty, condition uncertain, structural discrepancy, observational error and so forth.

We construct the implausibility measure as follows. Using the emulator we can obtain, for each set of inputs x, the mean and variance, $\mathrm{E}(F_h(x))$ and $\mathrm{Var}(F_h(x))$. If $x = x^*$, then , setting $F^* = F(x^*)$, we have

$$z_i = y_i \oplus e_i,\ y_i = F_i^* \oplus \epsilon_i$$

so that

$$\mathrm{Var}(z_i - \mathrm{E}(F_i(x))) = \mathrm{Var}(F_i(x)) + \mathrm{Var}(\epsilon_i) + \mathrm{Var}(e_i)$$

We can therefore calculate, for each output $F_i(x)$, the 'implausibility' if we consider the value x to be the best choice x^*, which is the standardized distance between z_i and $\mathrm{E}(F_i(x))$, given by

$$I_{(i)}(x) = |z_i - \mathrm{E}(F_i(x))|^2 / [\mathrm{Var}(F_i(x)) + \mathrm{Var}(\epsilon_i) + \mathrm{Var}(e_i)]$$

Large values of $I_{(i)}(x)$ suggest that it is 'implausible' that $x = x^*$.

The implausibility calculation can be performed univariately, or by multivariate calculation over sub-vectors for which we are prepared to make a full joint covariance specification for the emulator errors and for the structural discrepancy. With such a full joint specification, the implausibility criterion is a form of Mahalanobis distance between the system observations and the function outputs. The implausibilities are then combined, such as by using $I_M(x) = \max_i I_{(i)}(x)$, and can then be used to identify regions of x with large $I_M(x)$ as implausible, i.e., unlikely to be good choices for x^*.

Using this analysis, we can then refocus our efforts on the 'non-implausible' regions of the input space, by making more simulator runs and refitting our emulator over such sub-regions and repeating the analysis. This process is a form of iterative global search aimed at finding all choices of x^* which would give good fits to historical data.

We may find no choices at all which give good fits and that is a clear sign of problems with our physical simulator or with our data. Further, even if our ultimate goal is Bayesian model calibration, it is good practice to history match first, to check the model and (massively) reduce the search space for the Bayesian algorithm.

7 Forecasting

There are two basic sources of uncertainty that we must quantify in order to predict future system outcomes, y_p. Firstly, we are unsure as to the system prediction, $F_p(x^*)$, for y_p, as we are uncertain about both F and x^*, and secondly we are uncertain about the model discrepancy, ϵ_p, between $F_p(x^*)$ and y_p. The simplest Bayes linear forecasting system for taking account of these uncertainties is as follows; for details see [2].

The mean and variance of $F(x)$ are obtained from the mean function and variance function of the emulator f for F. Using these values, we compute the mean and variance of F^* by first conditioning on x^* and then integrating out x^*, typically over the parameter region identified by history matching. Given $\mathrm{E}(F^*), \mathrm{Var}(F^*)$, and specification of the variances for model discrepancy, ϵ, and sampling error, e, it is straightforward to compute the joint mean and variance of the collection (y, z) (as $y = F^* \oplus \epsilon, z = y_h \oplus e$).

We can therefore evaluate the mean and variance for y_p adjusted by z using the Bayes linear adjustment formulae. This analysis is fast and tractable even for large systems. Further, because of the simple structure of the calculations, it is tractable to identify collections of simulator evaluations which are appropriate for minimizing adjusted forecast variance. Typically, this will be the second stage choice of simulator evaluations, as the first stage will be a design appropriate to identify the form of emulator, estimate coefficient matrices and refocus, once or several times.

This analysis exploits the global features of the emulator to construct the joint covariance structure and is effective when the local component of emulator variation is small. When the local component is large, then a more detailed analysis is required, either by full Bayes specification or using the approach of Bayes linear calibrated forecasting; for details, see [7].

8 Example: emulating a climate simulator

(This uncertainty analysis is work with Danny Williamson, with NERC funding; details in [14].)

One of the aims of the NERC funded RAPID programme is to assess the risk of shutdown of the AMOC (Atlantic Meridionnal Overturning Circulation), which transports heat from the tropics to Northern Europe, and how this risk depends on the future emissions scenario for CO2. The RAPID sub-project aims to address aspects of this question by use of large ensembles of the UK Met Office climate model HadCM3, run through climate prediction.net. At an early stage of the project, as a preliminary demonstration of concept for the Met Office, we were asked to develop an emulator for HadCM3, based on 24 runs of the simulator, with a variety of parameter choices and future CO2 scenarios. We had access to some runs of FAMOUS (a lower resolution model), which consisted of 6 scenarios for future CO2 forcing, and between 40 and 80 runs of FAMOUS under each scenario, with different parameter choices. There was very little time to do the analysis.

The design that we chose was to match the inputs for 8 of the HadCM3 runs with corresponding inputs to a FAMOUS run (to help us to compare the models) and to construct a 16 run Latin hypercube over different parameter choices and CO2 scenarios (to extend the model across CO2 space). In this experiment only 3 parameters were varied (an entrainment coefficient in the model atmosphere, a vertical mixing parameter in the ocean, and the solar constant).

Our output of interest was a 170 year time series of AMOC values. The series is noisy and and the location and direction of spikes in the series was not important. Interest concerned aspects such as the value and location of the smoothed minimum of the series and the amount that AMOC responds to CO2 forcing and recovers if CO2 forcing is reduced.

To emulate the whole time series, we first smoothed by fitting splines $f^s(x,t) = \Sigma_j c_j(x) B_j(t)$ where $B_j(t)$ are basis functions over t and $c_j(x)$ are chosen to give the 'best' smooth fit to the time series. We emulate f^s by emulating each coefficient $c_j(x)$ in $f^s(x,t) = \Sigma_j c_j(x) B_j(t)$ (separately for each CO2 scenario). We test our approach by building emulators leaving out each observed run in turn, and checking whether the run falls within the stated uncertainty limits.

We now have an emulator for the smoothed version of FAMOUS, for each of the 6 CO2 scenarios. We extend the FAMOUS emulator across all choices of CO2 scenario using fast geometric arguments, exploiting the speed of working in inner product spaces. For example, we have a different covariance matrix for local variation at each of 6 CO2 scenarios. We extend this specification to all possible CO2 scenarios by identifying each covariance matrix as an element of an appropriate inner product space, and adjusting beliefs over covariance matrix space by projection.

We develop relationships between the elements of the emulator for FAMOUS and the corresponding emulator for HadCM3, using the paired runs, and expert judgements. This gives an informed prior for the HadCM3 emulator. We use the remaining runs of HadCM3 for Bayes linear adjustment of the emulator for HadCM3, and carry out further leave one out diagnostic checks and variance tuning. Our Met Office collaborators were happy with the resulting model emulations as a basis for further analysis given access to the larger ensemble.

9 Example: Oil Reservoir Simulators

(This uncertainty analysis is work with Jonathan Cumming, carried out with Basic Technology funding as part of the MUCM project; details of the application are in [4], and of the multi-level inference and design calculations are in [3].)

An oil reservoir is an underground region of porous rock which contains oil and/or gas. The hydrocarbons are trapped above by a layer of impermeable rock and below by a body of water, thus creating the reservoir. The oil and gas are pumped out of the reservoir and fluids are pumped into the reservoir (to boost production). The simulator models the flows and distributions of contents of the reservoir over time.

Each cell in the reservoir has a collection of associated input parameters, such as permeability and porosity. There are also other parameters, such as fault transmissibility, aquifer features and saturation properties. Since there are a huge number of cells in the reservoir, it is common to use scalar multipliers over subregions, to modify values.

The model outputs comprise the behaviour of the various wells and injectors in the reservoir Output, typically, is a time series on the following variables for each well; bottom-hole and tubing head pressure, production/injection rates and totals, for each of oil, water and gas, and fluid ratios for water cut and gas-oil ratio.

The term history matching, within the oil industry, refers to the identification of choices of input parameters for the simulator for which the simulator output is in close correspondence to the observed reservoir history. Our Bayes linear approach to reservoir history matching, based on the methodology described in [1], has been successfully implemented in software widely in use in the oil industry.

An example that we have provided to illustrate the methodology is given in [4]. This model, of a reservoir located in the North Sea, is based on grid size 38 × 87 × 25, with 43 production and 13 injection wells, and simulates 10 years of production, taking up to three hours per simulation. The inputs, in the illustration, are field multipliers for porosity (ϕ), permeabilities (k_x, k_z), critical saturation (crw), and aquifer properties (A_p, A_h). The outputs that we use for history matching are oil production rates for a 3-year period, for the 10 production wells active in that period, described by four month averages over the time series.

The computer model is expensive to evaluate, so we use a 'coarse' model, F^c, based on coarsening vertical gridding by factor of 10, to capture qualitative features of F. F^c is substantially faster, allowing 1000 runs of F^c in a Latin Hypercube over the input parameters.

Because of the high level of correlation between the different outputs, we use the principal variables approach to screen the wells. This method identifies, sequentially, the output, or group of outputs, which accounts for most of the variation in the remaining outputs. Applied to the coarse model evaluations, we retain outputs from 4 of the wells. These capture 87% of the total variation in all outputs.

We consider the coarse and the full model emulators to have the form

$$f_i^c(x) = \boldsymbol{g}_i(x_{[i]})^T \boldsymbol{\beta}_i^c + w_i^c(x), f_i(x) = \boldsymbol{g}_i(x_{[i]})^T \beta_i + w_i^c(x)\beta_{w_i} + w_i^a(x)$$

where $x_{[i]}$ is a subset of 'active inputs', i.e. the inputs which account for most of the variation in F. We fit emulators to each output individually, using stepwise regression and generalized least squares for the coarse model runs, to get emulator $f_i^c(x)$ for F_i^c. We found that the choice of three active inputs was adequate for expressing global variation in each output, for example achieving R^2 values in excess of 0.96 for all outputs but one. The porosity and critical saturation turned out to be active for all of the outputs, while each other output was active in a subset of the outputs. The two emulators are linked via equations relating corresponding pairs of coefficients as outlined in section (5). Careful choice of a small design of 20 evaluations for the full simulator, based on informative configuration over the active input collections, followed by Bayes linear adjustment, leads to the resulting emulator for F.

We now specify the observation and discrepancy variances and carry out the implausibility calculations for history matching. We find that working to a three standard deviation implausibility threshold eliminates about 90%of the input space, and corresponds to imposing a constraint on the upper value of ϕ. Since reducing the space, many of the old model runs are no longer relevant, so we supplement our emulation with further evaluations obeying the parameter constraint, namely an extra 100 coarse runs and 20 full simulator runs, and further adjust the emulator, using the old emulator structure as a starting point.

We now consider the final four time points in the three year period that we have emulated, and use the observed historical values to forecast the corresponding output values for an additional time point, one year beyond the end of this period. We have historical observations for the values to be forecast, which act as a quality check on the forecasts. We use the approach of section (7) effectively combining each model forecast with a correction for the estimated model discrepancy. In each case, the resulting forecast interval is within the measurement error of the actual historical measurement.

10 Example: Galaxy formation simulation

(This uncertainty analysis is work with Ian Vernon, carried out with with Basic Technology funding as part of the MUCM project; details in [13])

The Cosmologists at the Institute of Computational Cosmology at Durham University are interested in modelling galaxy formation in the presence of Dark Matter. First, a Dark Matter simulation is performed over a volume of (1.63 billion light years)^3. This takes 3 months on a supercomputer. Then, the simulator Galform takes the results of this simulation and models the evolution and attributes of approximately 1 million galaxies. Galform requires the specification of 17 unknown inputs in order to run. It takes approximately 1 day to complete 1 run (using a single processor).

The Galform model produces many outputs, some of which can be compared to observed data from the real Universe. Initially, we analyze luminosity functions giving the number of galaxies per unit volume, for each luminosity. These are Bj Luminosity, corresponding to density of young (blue) galaxies and K Luminosity, corresponding to density of old (red) galaxies. We choose 11 outputs that are representative of the Luminosity functions and emulate the functions $f_i(x)$.

We assess condition uncertainty, structural uncertainty, measurement uncertainty, and so forth. For example, we must account for the uncertainty resulting from the unknown configuration of dark matter in our universe. We can form judgements as to the magnitude of this uncertainty by making repeat simulations of Galform with the same input parameters and different choices of dark matter configuration.

We carry out the iterative history matching procedure, through four waves. For each wave, we evaluate the simulator many times, restricting parameter choices to those which have not yet been ruled out by earlier waves, emulate the simulator within the reduced space and carry out the implausibility calculations to reduce space further. A summary of the procedure, the number of active variables at each stage and the space removed at each stage is as follows.

	No. Model Runs	No. Active Vars	Space Remaining
Wave 1	1000	5	14.9 %
Wave 2	1414	8	5.9 %
Wave 3	1620	8	1.6 %
Wave 4	2011	10	0.12 %

In wave five, we evaluate many good fits to data, and we stop. Some of these choices give simultaneous matches to data sets that the Cosmologists have been unable to match before.

11 Linking models to reality

Each of the above examples, in common with most of the field of computer experiments, takes it as almost self-evident that the computer model is informative for the physical system. However, in most cases, the reason that the evaluations of the simulator are informative for the physical system is that the evaluations are informative about the general relationships between system properties, x, and system behaviour y. Therefore, our inference from model to reality should proceed in two parts.

We emulate the relationship between system properties and system behaviour. We call this relationship, F^*, the “reified model” (from reify: to treat an abstract concept as if it were real). We can then decompose the difference between our model and the physical system into two parts. The first is the difference between our simulator and the reified form, and the second is the difference between the reified form at the physically appropriate choice of x and the actual system behaviour y. We call this the "Reifying principle", namely that the simulator F is informative for y, because F is informative for F^* and $F^*(x^*)$ is informative for y. Similarly, a collection of simulators $F_1, F_2, ...$ is jointly informative for y, as the simulators are jointly informative for F^*.

We link F and F^* using emulators. Suppose that our emulator for F is

$$f(x) = Bg(x) \oplus u(x)$$

Our simplest emulator for F^* might be

$$f^*(x, w) = B^* g(x) \oplus u^*(x) \oplus u^*(x, w)$$

where we might model our judgements as $B^* = CB + \Gamma$ and correlate $u(x)$ and $u^*(x)$, while treating $u^*(x, w)$, with additional parameters, w, as uncorrelated with the remaining terms in the emulator. Structured reification improves on this with systematic modelling for all aspects of model deficiency whose effects we can consider explicitly. For an illustrated treatment of reification, see [8].

All of the Bayes linear history matching and forecasting methodology that we have described is unchanged by this extra layer of modelling. All that has changed is our description of the joint covariance structure which underlies each of the subsequent calculations.

12 Concluding comments

To assess our uncertainty about complex systems, it is enormously helpful to have an overall (Bayesian) framework to unify all of the sources of uncertainty. Within this framework, all of the scientific, technical, computational, statistical and foundational issues can be addressed in principle. Such analysis poses serious challenges, but they are no harder than all of the other modelling, computational and observational challenges involved with studying complex systems.

In particular, Bayes and Bayes linear multivariate, multi-level, multi-model emulation, careful structural discrepancy modelling and iterative history matching gives a great first pass treatment for most large modelling problems.

References

1. Craig P.S., Goldstein M., Seheult A.H., Smith J.A. (1997): Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments (with discussion), in Case Studies in Bayesian Statistics, vol. III, eds. C. Gastonis et al. 37-93. Springer-Verlag.
2. Craig P.S., Goldstein M., Rougier J.C., and Seheult, A.H.: Bayesian Forecasting for Complex Systems Using Computer Simulations. Journal of the American Statistical Association 96, 717-729 (2001)

3. Cumming, J. and Goldstein, M.: Small Sample Bayesian Designs for Complex High-Dimensional Models Based on Information Gained Using Fast Approximations. Technometrics, 51 377-388 (2009)
4. Cumming, J. and Goldstein, M.: Bayes Linear Uncertainty Analysis for Oil Reservoirs Based on Multiscale Computer Experiments. In The Oxford Handbook of Applied Bayesian Analysis. O'Hagan, and West,A.M. Oxford University Press. 241-270 (2009)
5. de Finetti, B.: Theory of Probability, vol 1 & 2, Wiley (1974,1975)
6. Goldstein, M.: Subjective Bayesian analysis: principles and practice. Bayesian Analysis 1: 403-420 (2006)
7. Goldstein, M. and Rougier, J.C.: Bayes linear calibrated prediction for complex systems. Journal of the American Statistical Society, 101, 1132-1143 (2006)
8. Goldstein, M. and Rougier, J.C.: Reified Bayesian modelling and inference for physical systems (with discussion), Journal of Statistical Planning and Inference, 139, , 1221-1239 (2008)
9. Goldstein, M. and Wooff, D.A.: Bayes Linear Statistics: Theory and Methods, Wiley (2007)
10. Kennedy, M.C. and O'Hagan, A.: Bayesian calibration of computer models (with discussion). Journal of the Royal Statistical Society, B,63, 425-464 (2001)
11. O'Hagan, A.: Bayesian analysis of computer code outputs: a tutorial. Reliability Engineering and System Safety 91 (2006)
12. Santner, T., Williams, B. and Notz, W.: The Design and Analysis of Computer Experiments. Springer Verlag: New York (2003)
13. Vernon I., Goldstein M., and Bower, R.: Galaxy Formation: a Bayesian Uncertainty Analysis (with discussion). Bayesian Analysis, 5, 619-670 (2010)
14. Williamson, D. and Goldstein, M: Fast linked analyses for scenario based hierarchies. Journal of the Royal Statistical Society: Series C , to appear (2012)

DISCUSSION

Speaker: Michael Goldstein

Kyle Hickmann : Could you speak a little bit more on in what sense the emulator converges to the simulator as more points are observed?

Michael Goldstein : The emulator is exactly equal to the simulator at each observation point, and uncertainty about the simulator increases for input choices far from any of the observed values. We reduce uncertainty in any region of parameter space by making function evaluations in that region, and the more evaluations that we make, the further will the uncertainty be reduced. How large a sample we must make to achieve a good measure of convergence across the whole input space depends on the dimension of the input and output spaces and the degree of regularity of the function over the range of the input space. For example, very small regions of input space, in which the function behaves quite differently from behaviour everywhere else, can be extremely difficult to identify and emulate appropriately.

Antonio Possolo : We have learned from Lindley that linear polling is one way of merging the conclusions multiple Bayesian analyses will have produced. What is the state of the art?

Michael Goldstein : The appropriate way to merge multiple Bayesian analyses depends on your judgements about the level of, and the relationship between, the information and expertise contained within each analysis. There is no automatic way to do this. The reification formalism described in this article is one way of structuring the joint analysis when dealing with Bayesian analyses based around computer simulators.

Antonio Possolo : An analysis that starts from expectations, variances, and covariances, is bound to produces results that are expectations, variances, and covariances. What additional assumptions would you regard as defensible to be able to quantify the conclusions probabilistically?

Michael Goldstein : Probabilities are themselves expectations, for the indicator functions corresponding to the events. If the analysis is described at a sufficient level of detail to identify some of these expectations, then we have a direct probabilistic inference. Alternately, we can use qualitative probabilistic judgements to make a low assumption bridge between the Bayes linear and the full probabilistic analysis. For example, when carrying out a history matching analysis as described in this article, it is useful to know that, for any continuous, unimodal probability density function, 95% of the probability will be contained within three standard deviations of the mean (the so-called 3 sigma rule).

Jeffrey Fong : Regarding Bayesian linear analysis, in your two equations, one expectation and the second variance, do they allow a user to derive a host of relationship (as in classical theory of error propagation) that are used to get expectation and variance of a sum, product, quotient, etc... of a complicated algebraic form?

Michael Goldstein : Bayes linear inferences obey all of the rules derived from the linearity of expectation. Therefore, it is necessary to ensure that the appropriate polynomial or other functional forms of the quantities of interest are introduced as elements of the adjusting vector and of the vector of terms to be adjusted. The Bayes linear Statistics volume ([9]) contains examples and discussions of this.